INDONESIAN GRAPHEME-TO-PHONEME (G2P) MENGGUNAKAN MODEL IG-TREE + STRATEGI TEBAKAN TERBAIK

Agus Hartoyo[1], Suyanto[2], Dyas Puspandari S.s. M.pd[3]

[1]Teknik Informatika, Fakultas Teknik Informatika, Universitas Telkom

---

Abstrak

Konversi Indonesian grapheme-to-phoneme (G2P) merepresentasikan sebuah tugas memetakan setiap grafem / symbol eja dalam sembarang kata yang dikenal dalam bahasa Indonesia ke representasi fonemik / simbol pelafalannya.
Pencarian metode terbaik yang dilakukan dalam tugas akhir ini memberikan hasil berupa penetapan sebuah model bernama IG-tree + strategi tebakan-terbaik sebagai metode yang terpilih untuk memecahkan permasalahan konversi G2P. Model tersebut pada dasarnya menggunakan struktur pohon-keputusan yang dibangun berdasarkan data training, dikonstruksikan menggunakan konsep information gain (IG) dalam menentukan kepentingan relatif atribut-atribut, dan dilengkapi dengan strategi tebakan-terbaik dalam mengklasifikasikan instan-instan baru. Akan tetapi sistem dalam tugas akhir ini dikembangkan lebih lanjut dengan properti-properti baru yang ditambahkan pada struktur asalnya untuk meningkatkan performansi sistem. Mekanisme pruning diusulkan untuk model dengan dua tujuan: (1) meningkatkan kemampuan generalisasi model, dan (2) meminimalkan ukuran model. Properti baru yang lain, peng-handle kasus homograf menggunakan metode kategorisasi teks, diusulkan untuk sistem untuk menangani kasus khususnya berupa beberapa himpunan kata yang sepenuhnya sama dalam representasi grafemik namun berbeda satu sama lain dalam representasi fonemik. Ditunjukkan dalam tugas akhir ini bahwa model tersebut secara umum berperformansi bagus sementara properti-properti tambahan yang diusulkan memang memberikan keuntungan tambahan sebagaimana yang diharapkan.

Kata Kunci : konversi grapheme-to-phoneme, bahasa Indonesia, IG-tree, strategi tebakan-terbaik, pruning, peng-handle kasus homograf

---

Abstract

Indonesian grapheme-to-phoneme (G2P) conversion represents a task of mapping each grapheme / spelling symbol in any Indonesian word to its phonemic representation / pronounciation symbol A selection for the best method is in this final project results in determining a model called IG-tree + best-guess strategy as the chosen model to solve G2P conversion problem. The model is basically in decision-tree structure built based on a trainingset, constructed using concept of information gain (IG) in weighing the relative importance of attributes, and equipped with the best-guess strategy in classifying new instances. However, the system is in this final project leveraged with new features added to its pre-existing structure to improve its performance. A pruning mechanism is proposed for the model for two objectives: (1) improving its generalization ability, and (2) minimizing its dimension. Another new feature, the homograph case handler using a text-categorization method, is proposed for the system to handle its special case of a few sets of words which are exactly the same in graphemic representations but are different each other in phonemic representations.
It is shown in this final project that the model in general performs well while the additional features really give additional benefits as expected.

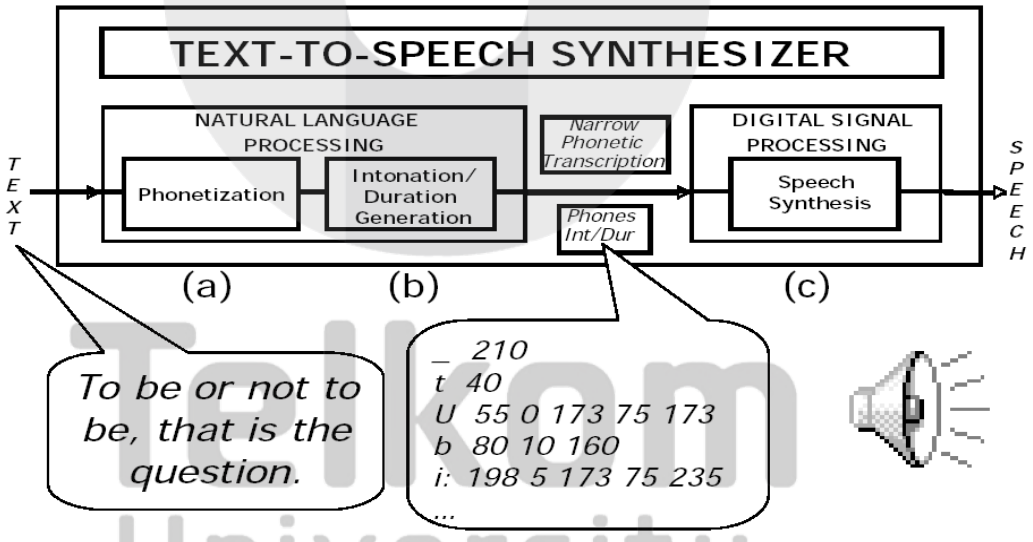
Keywords : grapheme-to-phoneme conversion, Indonesian, IG-tree, best-guess strategy, pruning, homograph-case handler

---

1. Pendahuluan

1.1 Latar belakang masalah

Konversi grafem-ke-fonem (*grapheme-to-phoneme* / G2P) merupakan salah satu modul utama dalam sistem *text-to-speech* (TTS). Posisi modul ini dalam sistem TTS digambarkan di dalam Gambar 1-1 sebagai sistem berindeks (a) bernama Phonetization. Diberikan sebuah abjad simbol-simbol eja (grafem-grafem) dan abjad simbol-simbol pelafalan (fonem-fonem), sebuah pemetaan harus dapat mentransliterasikan rangkaian-rangkaian grafem ke rangkaian-rangkaian fonem mereka [5]. Pendekatan berbasis aturan untuk sistem konversi *G2P*, yang berarti membebaskan sistem dari ketergantungannya pada kelengkapan leksikon, sangat dibutuhkan mengingat bahwa kosa kata dalam sebuah bahasa akan terus berkembang dengan waktu yang memungkinkan lahirnya kata-kata baru yang tidak ada pada leksikon ”lengkap” yang disusun sebelum kata-kata baru tersebut muncul (Sebagai contoh, sampai hari ini di dalam Kamus Besar Bahasa Indonesia tidak tercantum kata *tetikus* dan kata *unduh* yang baru muncul belakangan ini untuk menggantikan berturut-turut kata *mouse* dan kata *download*). Di samping itu kebutuhan akan pendekatan berbasis aturan untuk sistem konversi G2P juga terkait dengan adanya kasus-kasus di mana kapasitas memori sistem sangat terbatas. Dalam kasus-kasus demikian penggunaan leksikon berukuran besar tentu saja sangat tidak praktis.



[T. Dutoit 2002 – Faculty Polytechnique de Mons Belgium]

Gambar 1-1 Skema modul-modul penyusun sistem *text-to-speech*

Mengingat bahwa bahasa yang berbeda memiliki aturan grafemik dan fonetik serta morfologi yang berbeda, idealnya sebuah sistem konversi G2P dibuat spesifik untuk sebuah bahasa tertentu. Meskipun sebuah sistem konversi

G2P yang berkategori bebas bahasa (*language independent)* secara umum adalah dapat digunaulang (*reusable*) untuk bahasa manapun, sentuhan-sentuhan khusus yang khas sebuah bahasa tetap dibutuhkan untuk memberikan performansi terbaik bagi sistem pada bahasa yang bersangkutan. Dalam hal ini Indonesian G2P merupakan sistem konversi G2P yang dinisbatkan secara khusus pada bahasa Indonesia dan merepresentasikan sebuah mesin yang memetakan setiap grafem penyusun sebarang kata yang dikenal dalam bahasa Indonesia ke fonemnya.

Gambar 1-2 memperlihatkan secara kotak hitam apa yang dilakukan oleh sistem Indonesian G2P.

Gambar 1-2 Sistem Indonesian G2P *dengan masukan dan keluarannya*

Berbagai teknik dan metode / model telah diusulkan untuk mengimplementasikan sistem konversi G2P berbasis aturan. Metode yang paling tradisional adalah metode berbasis-pengetahuan kebahasaan (*linguistic knowledge-based method*) yang berusaha memformalisasikan berbagai pengetahuan kebahasaan (meliputi tata bahasa, aturan pemenggalan suku kata, morfologi, fonologi, dsb.) beserta interaksi di antara sumber-sumber kebahasaan itu menjadi aturan-aturan, kemudian menggunakan aturan-aturan itu untuk mengkorversi sebarang kata menjadi representasi fonemiknya. Pendekatan ini dianggap kurang praktis karena 2 alasan: (1) pengetahuan yang dibutuhkan di dalamnya sangat bergantung pada bahasa; dan (2) pendekatan ini membutuhkan begitu banyak rekayasa kebahasaan secara eksplisit. Untuk menjawab kedua masalah pada pendekatan terikat-bahasa (*language-dependent approach*) tersebut Antal van den Bosch dan Walter Daelemans dengan *instance-based learning* serta IG-Tree dengan strategi tebakan terbaik mereka, Francois Yvon dengan *chunk-based pronounciation*-nya, serta Paul Taylor dengan *hidden Markov model*-nya, mengusulkan pendekatan yang bersifat bebas bahasa. Selain bahwa sistem yang dihasilkan dapat digunaulang pada berbagai *data set* dari bahasa-bahasa yang berbeda, keunggulan yang juga dimiliki oleh pendekatan ini adalah bahwa sistem yang dihasilkan bersifat otomatis di mana rekayasa kebahasaan secara eksplisit tidak lagi dibutuhkan. Berbicara tentang performansi, uji coba-uji coba yang telah dilakukan menunjukkan bahwa hasil yang dapat dicapai melalui pendekatan bebas-bahasa ternyata sama atau bahkan lebih baik dibanding yang dapat dicapai melalui pendekatan terikat-bahasa, yakni sebagaimana yang telah diperlihatkan oleh Antal van den Bosch dan Walter Daelemans dalam [4].

Pemilihan pendekatan bebas-bahasa sebagai pendekatan yang digunakan dalam tugas akhir ini sesungguhnya didasarkan pada sejumlah keunggulan dan kebaikannya yang telah disebutkan di atas. Dengan dasar itu pula pemilihan model yang akan digunakan dalam tugas akhir ini akan dilakukan – dengan dasar aspek keunggulan dan kebaikannya. Keunggulan dan kebaikan sebuah model akan dinilai dari berbagai sisi dalam tugas akhir ini melalui empat parameter:

1. Tingkat kecepatan
   Tingkat kecepatan sebuah model dapat dilihat dari kompleksitas algoritma yang dipekerjakan di dalam model itu. Kecepatan berbanding terbalik dengan kompleksitas algoritma.
2. Tingkat kehematan-memori
   Tingkat kehematan-memori adalah ukuran positif dari tingkat kebutuhan-memori yang negatif. Tingkat kehematan memori ditentukan oleh bentuk aturan yang dihasilkan setelah proses pelatihan dan bagaimana cara sistem menyimpan aturan tersebut.
3. Akurasi per fonem
   Akurasi per fonem adalah banyaknya fonem benar yang berhasil dipetakan dibagi dengan total banyaknya seluruh fonem dalam *test set*. Akurasi per fonem dinyatakan dalam persen
4. Akurasi per kata
   Akurasi per kata adalah banyaknya kata yang setiap grafemnya dipetakan menjadi fonem yang benar dibagi dengan total banyaknya seluruh kata dalam *test set*. Akurasi per kata dinyatakan dalam persen

Tabel 1-1 menunjukkan bagaimana parameter-parameter di atas diukurkan pada beberapa model yang menggunakan pendekatan bebas-bahasa, dikuantitatifkan, kemudian dihitung menjadi skor-skor. Nilai-nilai pada parameter akurasi per fonem dan akurasi per kata untuk masing-masing model diperoleh dari hasil uji coba yang telah dilakukan oleh para peneliti dan dipublikasikan di dalam [4], [5], [6], [11], dan [12], dalam tugas akhir ini dengan mengabaikan *data set* dan bahasa yang berbeda-beda yang digunakan oleh masing-masing peneliti. Baik parameter tingkat kecepatan maupun tingkat kehematan memori ditetapkan memiliki domain nilai 80 untuk kualifikasi rendah, 90 untuk kualifikasi sedang, dan 100 untuk kualifikasi tinggi. Di sini domain nilai pada kedua parameter tersebut ditetapkan memiliki kisaran angka dari 80 hingga 100 dengan tujuan untuk menyeimbangkan pengaruh kedua parameter (pada proses penghitungan skor) dengan paramater-parameter akurasi yang juga memiliki kisaran nilai pada angka-angka itu. Skor dirumuskan sebagai 35% tingkat kecepatan + 15% tingkat kehematan memori + 25% akurasi per fonem + 25% akurasi per kata. Kebijaksanaan pembobotan di sini ditetapkan untuk maksud-maksud praktis. Kedua parameter akurasi bersama-sama diberi bobot terbesar, yakni total 50%, karena akurasi adalah aspek yang paling mengaktualisasikan maksud dibuatnya sebuah sistem konversi G2P. Tingkat kecepatan memiliki bobot 35% – bobot yang cukup signifikan – karena berbicara tentang kecepatan adalah berbicara tentang *delay*, dan masalah *delay* adalah masalah yang sangat kritis untuk beberapa sistem tertentu yang mengaplikasikan sistem konversi G2P seperti sistem *Speech-to-Speech Machine Translation* (S2SMT) di mana *delay* sekecil apapun memiliki pengaruh signifikan pada performansi sistem. Selanjutnya bobot terkecil diberikan kepada tingkat

kehematan-memori dengan mengingat hal-hal berikut ini: (1) format data yang biasanya harus disimpan dalam memori penyimpanan sebuah sistem konversi G2P adalah format teks yang notabene memiliki ukuran yang tidak signifikan; (2) pada sistem S2SMT, misalnya, toh sistem konversi G2P yang digunakan – termasuk memori penyimpanannya – ditempatkan pada server operator; dan (3) kalaupun pada pengembangan lebih lanjut sistem konversi G2P diarahkan untuk ditanamkan pada pesawat telepon seluler, data yang harus disimpan yang sesungguhnya tidak akan lebih dari 1 Mbyte tidak akan menjadi masalah bagi pesawat hari ini.

Model	Tingkat Kecepatan	Tingkat Kehematan Memori	Akurasi per Fonem	Akurasi per Kata	Skor
TABLE LOOKUP + FIXED DEFAULTS	90	80	95,10	89,50	89,65
IG-TREE + STRATEGI TEBAKAN TERBAIK	90	90	95,10	89,50	91,15
IG-TREE + *SBR*	90	90	97,40	83,70	90,28
ITERATIVE DICHOTOMIZER 3	90	80	90,00	60,00	81,00
HIDDEN MARKOV MODEL + STRESS ADJUSTMENT	100	100	92,28	61,08	88,34

Tabel 1-1: Perbandingan beberapa model konversi G2P yang menggunakan *pendekatan bebas-bahasa*

Atribut Skor pada Tabel 1-1 merupakan besaran yang menunjukkan tingkat keunggulan dan kebaikan model yang bersangkutan, dan karena itu, menjadi besaran yang akan secara langsung dibandingkan untuk memilih model. Terlihat pada tabel bahwa model IG-Tree + strategi tebakan terbaik menempati urutan teratas dalam skor dengan angka 91,15. Hal itu menyarankan agar model IG-Tree + strategi tebakan terbaik dipilih sebagai model yang digunakan untuk membangun sistem Indonesian G2P dalam tugas akhir ini.

Meskipun secara *overall* model yang dipilih paling unggul di antara model-model yang ada, model tersebut masih memiliki beberapa aspek yang bisa dikembangkan. Dengan melihat fakta bahwa model tersebut dibangun dalam struktur pohon-keputusan yang mengkompres data secara *lossless*, penulis mengusulkan sebuah pengembangan pada aspek itu dengan menawarkan diterapkannya mekanisme *pruning* (pemangkasan) atas model. Dengan mekanisme *pruning* penyimpanan data oleh model tidak lagi *lossless*, melainkan – diharapkan – *lossy* pada aturan-aturan yang mengakomodasi *outlier* sehingga kemampuan generalisasi model meningkat. Di samping itu satu keuntungan lain yang jelas akan didapatkan dengan menerapkan mekanisme tersebut adalah bahwa dimensi model akan menjadi lebih kecil.

Lebih jauh lagi sesungguhnya terdapat sebuah permasalahan permanen dalam topik konversi G2P yang tidak pernah secara khusus ditangani oleh satupun metode di atas. Mekanisme pemeriksaan konteks berbasis huruf yang diterapkan di dalam metode-metode tersebut yang hanya melibatkan huruf-huruf konteks dalam internal kata, tidak memungkinkan mereka untuk dapat secara memuaskan memecahkan masalah homograf. Dengan mekanisme yang semacam itu, mereka

tidak akan pernah dapat secara beralasan menentukan jawaban untuk pertanyaan ”Apakah gerangan representasi fonemik dari grafem <e> pada kata *apel*?” selain dari hanyalah penebakan acak. Dalam kasus ini, bahkan manusia – sebuah sistem ideal yang hendak ditiru oleh sistem konversi G2P – pun tidak akan berdaya menghadapi pertanyaan yang sama tanpa diberi tahu (atau mencari tahu) konteks yang lebih luas di sekitar kata homograf itu.

Untuk dapat memecahkan permasalahan homograf, sebuah sistem konversi G2P haruslah menerapkan mekanisme pemeriksaan konteks yang lebih luas – dalam contoh kasus di atas, yang menyarankan sistem untuk ”mencari tahu” kalimat macam apa sesungguhnya yang melingkupi kata *apel*, apakah itu semacam kalimat *Buah apel biasanya merah di luar saat masak (siap dimakan), namun bisa juga hijau atau kuning* ataukah semacam kalimat *Seribu anggota Garda Bangsa di Jawa Timur menggelar apel pagi bersama di lapangan parkir Jatim Expo, Surabaya, hari ini*. Sangat jelas bahwa permasalahan ini tidak lain dari permasalahan klasifikasi/kategorisasi teks, sebuah permasalahan utama dalam *text mining* yang berupaya menentukan kategori/topik sebuah teks dengan memperhatikan *term-term* yang dikandung oleh teks itu.

Kosa kata homograf hanyalah merupakan bagian yang sangat kecil dari keseluruhan kosa kata dalam sebuah bahasa, tapi kemampuan memecahkan permasalahan homograf tetap merupakan nilai tambah bagi sistem konversi G2P yang memiliki kemampuan demikian. Dengan menerapkan salah satu metode kategorisasi teks, sistem Indonesian G2P yang akan dibangun dalam tugas akhir ini dirancang untuk memiliki kemampuan tersebut .

Oleh sebab permasalahan homograf hanyalah merupakan sebuah sub-permasalahan yang khusus dari permasalahannya konversi *G2P* secara umum, masalah pemilihan metode penanganan homograf – dalam hal ini masalah pemilihan metode kategori teks yang tepat – tentu saja tidak sekrusial masalah penentuan metode utama konversi *G2P* sebelumnya. Cukuplah dengan berargumen pada pernyataan ”Eksperimen ekstensif kami menunjukkan bahwa *classifier* berbasis centroid secara konsisten dan substansial berperformansi melebihi algoritama-algoritma lain seperti Naive Bayes, k-nearest-neighbor, dan C45, pada dataset-dataset yang beraneka.” [7], penulis memilih metode berbasis centroid sebagai metode kategori teks yang akan digunakan untuk memecahkan permasalahan homograf dalam tugas akhir ini.

1.2 Perumusan Masalah

Berdasarkan latar belakang masalah yang dikemukakan di atas penulis merumuskan bahwa masalah-masalah yang akan diselesaikan dengan riset tugas akhir ini adalah sebagai berikut:

1. bagaimana mengimplementasikan model *IG-Tree* + strategi tebakan terbaik untuk membangun sistem *Indonesian G2P* secara umum;
2. bagaimana mengimplementasikan properti baru yang diusulkan berupa diterapkannya mekanisme *pruning* pada pembangunan model *IG-tree*;
3. bagaimana mengimplementasikan properti baru yang diusulkan berupa diterapkannya metode kategorisasi teks berbasis centroid untuk memecahkan permasalahan homograf dalam sistem *Indonesian G2P*; dan

4. bagaimana mengukur dan menganalisis performansi sistem Indonesian G2P yang dibangun secara umum;
5. bagaimana mengukur dan menganalisis pengaruh properti-properti baru pada performansi sistem.

Dalam rangka memecahkan kelima masalah tersebut penulis menetapkan batasan bahwa *dataset* yang digunakan baik untuk melatih sistem Indonesian G2P maupun untuk mengukur performansinya adalah leksikon pasangan pelafalan-kata yang berisi *hanya* kata-kata yang dikenal dalam bahasa Indonesia, termasuk di dalamnya kata-kata serapan atau kata-kata asing yang sudah diindonesiakan. Dengan demikian leksikon yang mengandung kata asing (yang belum diindonesiakan) akan dikategorikan dalam riset ini sebagai *dataset* yang tidak valid.

1.3 Tujuan

Berdasarkan rumusan masalah yang dikemukakan di atas penulis menetapkan tujuan riset tugas akhir ini sebagai berikut:

1. menerapkan model IG-Tree + strategi tebakan terbaik untuk membangun sistem Indonesian G2P secara umum;
2. menerapkan mekanisme *pruning* dalam pembangunan model *IG-tree* sebagai properti tambahan yang diusulkan oleh penulis;
3. menerapkan metode kategorisasi teks berbasis centroid untuk memecahkan permasalahan homograf dalam sistem *Indonesian G2P* sebagai properti tambahan yang diusulkan oleh penulis;
4. menguraikan analisis tentang hasil pengukuran performansi sistem Indonesian G2P yang diimplementasikan;
5. menguraikan analisis tentang pengaruh properti-properti tambahan terhadap performansi system.

1.4 Metodologi Penyelesaian Masalah

Metode yang akan digunakan untuk menyelaikan permasalahan dalam tugas akhir ini adalah sebagai berikut:

1. Studi pustaka
   Penulis mempelajari dasar teori tentang pendekatan-pendekatan dan model-model yang diusulkan dalam masalah G2P, terutama model *IG-Tree* + strategi tebakan terbaik, melalui makalah-makalah yang dipublikasikan melalui internet oleh para peneliti. Penulis juga membaca pustaka-pustaka yang membahas teori-teori kebahasaIndonesiaan, khususnya yang berkaitan dengan fonologi..
2. Pembangunan leksikon
   Fase 1, penulis mengambil 50.000 – 100.000 kata mentah dari surat kabar *online*.
   Fase 2, penulis melakukan operasi *distinct* atas himpunan kata tersebut.
   Fase 3, penulis menyortir korpus outputan fase 2 hingga menjadi 6.000 – 7.000 kata saja yang benar-benar representatif.
   Fase 4, untuk setiap kata outputan fase 3 penulis memberi setiap grafem penyusun kata itu dengan pasangan fonemnya secara manual.

Outputan dari fase 5 adalah leksikon pasangan pelafalan-kata yang akan berperan sebagai *dataset* untuk sistem Indonesian G2P.

3. Validasi *dataset* oleh ahli bahasa Indonesia.
4. Pembagian *dataset* menjadi *training set, validation set,* dan *test set.*
5. Perancangan perangkat lunak Indonesian G2P menggunakan UML.
6. Implementasi perangkat lunak Indonesian G2P menggunakan bahasa Java dengan editor Netbeans 5.5.
7. Pelatihan sistem Indonesian G2P dengan *training set* untuk membangun *IG-Tree*, dan *validation set* untuk melakukan *Pruning* validasi pada *IG-Tree.*
8. Pengujian sistem Indonesian G2P dengan *test set.*
9. Analisis atas hasil pengujian sistem Indonesian G2P
10. Pengambilan kesimpulan dan penulisan laporan.

5. Penutup

5.1 Kesimpulan

Setelah menjalankan langkah-langkah perancangan, implementasi, dan analisis pada bab-bab terdahulu, penulis berhasil merumuskan poin-poin kesimpulan sebagai berikut:

1. Sistem pembangunan model *IG-Tree* + strategi tebakan terbaik dapat diimplementasikan untuk memecahkan permasalahan konversi *Indonesian G2P*.
2. Sistem pembangunan model *IG-tree* + strategi tebakan terbaik memiliki sejumlah properti dan karakteristik yang mendukungnya menjadi sistem yang sangkil dan mangkus dalam memecahkan permasalahan konversi *Indonesian G2P*.
   a. Model yang dihasilkan yang memiliki struktur dasar *tree* memiliki karakteristik dasar sebagai kompresor data yang *loosless* dan sebagai pengekstrak data menjadi aturan.
   b. Sistem pembangunan model *IG-tree* + strategi tebakan terbaik memiliki properti pemetaan satu-ke-satu. Properti ini sangat bermanfaat dalam meminimalkan ambiguitas dan menyederhanakan algoritma.
   c. Sistem pembangunan model *IG-tree* + strategi tebakan terbaik memiliki properti pengurutan besarnya nilai *information gain* konteks untuk menentukan urutan perluasan konteks. Properti ini mendukung sistem menjadi sistem yang sangkil dalam fase pembangunan model serta *classifier*, dan mangkus dalam kebutuhan ruang.
   d. Sistem pembangunan model *IG-tree* + strategi tebakan terbaik memiliki properti tambahan berupa mekanisme *pruning*. Properti ini secara meyakinkan selalu berhasil menekan ukuran model meskipun kemampuannya untuk menghasilkan model yang paling representatif sangat ditentukan oleh tingkat representativitas *dataset*.
   e. Sesuai dengan namanya sistem pembangunan model *IG-tree* + strategi tebakan terbaik memiliki properti strategi tebakan terbaik yang berhasil memaksimalkan aspek generalisasi aturan dalam *classifier*.
3. Pada level *phoneme-based* model akhir yang dihasilkan memiliki performansi yang tinggi dan sangat bisa diterima yang ditunjukkan dengan nilai akurasi per fonem 99.01% dan nilai akurasi per kata 92.49%.
4. Pada level *phoneme-based* mekanisme *pruning* berhasil menekan ukuran model sehingga ukuran model menjadi minimum dan sangat bisa diterima. Hal itu ditunjukkan oleh ukuran file penyimpan model (teks) akhir yang hanya membutuhkan ruang sebesar 8.60 KB saja,
5. Pada level *alophone-based* model akhir yang dihasilkan memiliki performansi yang tidak terlalu baik tapi masih bisa diterima yang ditunjukkan dengan nilai akurasi per fonem 95.26 %dan nilai akurasi per kata 67.30%. Diindikasikan mekanisme *pruning* dengan *validation set* yang kurang contoh hanya meningkatkan akurasi validasi secara gradual tapi pada saat yang sama justeru juga secara gradual menurunkan akurasi testing.
6. Lebih jauh diindikasikan bahwa keseluruhan *dataset* berbasis alofon yang digunakan pada tugas akhir ini secara umum menghadapi permasalahan

kurang contoh di hadapan tugas pembangunan aturan pemetaan yang cukup rumit pada level ini.

7. Meski begitu pada level *alophone-based* mekanisme *pruning* paling tidak telah berhasil menekan ukuran model sehingga ukuran model menjadi minimum dan sangat bisa diterima. Hal itu ditunjukkan oleh ukuran file penyimpan model (teks) akhir yang hanya membutuhkan ruang sebesar 18.3 KB saja.
8. Secara umum baik pada level *phoneme-based* maupun *alophone-based* mekanisme pruning telah berhasil menekan dimensi model secara signifikan. Hal itu ditunjukkan dengan rasio antara dimensi model setelah di-*prune* dengan dimensi model sebelum di-*prune* yang berada di sekitar angka 2/5.
9. Pada saat yang sama, pengurangan dimensi model yang signifikan selama proses *pruning* tersebut tidak berakibat pada jatuhnya kemampuan *classifier*. Hal itu ditunjukkan oleh nilai akurasi model pada level *phoneme-based* yang tetap sangat tinggi pasca-*pruning*.
10. Pendekatan kategorisasi teks / *text mining* yang ditawarkan oleh penulis dapat secara sempurna memecahkan permasalahan homograf dengan syarat (i) konteks kalimat dalam teks-teks *training set* cukup representatif, dan (ii) konteks kalimat dalam teks-teks *test set* benar-benar mengarah ke topik / kategori dan tidak mengecoh.

5.2 Saran

Dengan memperhatikan kendala-kendala yang dihadapi penulis selama pembangunan sistem dan dengan mempertimbangkan kekurangan-kekurangan dari sistem yang telah dibangun, penulis mengajukan saran-saran sebagai berikut:

1. *Dataset* perlu dikembangkan agar performansi model berbasis alofon meningkat.
2. Perlu diujicobakan pengembangan yang lebih lanjut dari sistem pembangunan model ini yang menghitung ulang *information gain* setiap kali memperluas konteks. Perlu dilihat apakah peningkatan kompleksitas algoritma yang ditanggung sistem sepadan dengan peningkatan performansi model yang dihasilkan atau tidak.
3. Sekat-sekat yang tidak perlu antardisiplin ilmu harus dihilangkan mengingat bahwa terdapat begitu banyak permasalahan di dunia riil yang membutuhkan kerjasama dan sinergi lintasdisiplin ilmu.

Daftar Pustaka

[1] Alwi, Hasan et al. 2003. *Tata Bahasa Baku Bahasa Indonesia*. Edisi Ketiga. Jakarta: Balai Pustaka.

[2] Asia, Jelita et al. Stemming Indonesian. 2006. http://portal.acm.org/ft_gateway.cfm?id=1082195&type=pdf&coll=GUIDE&dl=ACM&CFID=20706954&CFTOKEN=66370341., didownload pada tanggal 1 Maret 2007.

[2] Bouma, Gosse. A Finite State and Data-Oriented Method for Grapheme-to-Phoneme Conversion. 2000. http://www.let.rug.nl/~gosse/papers/naacl.ps , didownload pada tanggal 1 Maret 2007.

[3] Caseiro, D, et al. Grapheme-to-Phoneme Using Finite-State Transducer. http://www.inesc-id.pt/pt/indicadores/Ficheiros/183.pdf didownload pada tanggal 1 Maret 2007.

[4] Daelemans, Walter dan Antal van den Bosch. Tabtalk: Reusability in Data-Oriented Grapheme-to-Phoneme Conversion. 1993 http://www.cnts.ua.ac.be/Publications/1993/DV93/db93.pdf didownload pada tanggal 1 Maret 2007 .

[5] ——. Data-Oriented Method for Grapheme-to-Phoneme Conversion. 1993. http://acl.ldc.upenn.edu/E/E93/E93-1007.pdf didownload pada tanggal 1 Maret 2007

[6] —— . Language-Independent Data-Oriented Grapheme-to-Phoneme Conversion. 1997. http://www.cnts.ua.ac.be/~walter/papers/1996/db96.ps didownload pada tanggal 1 Maret 2007.

[7] Han, Eui-Hong dan George Karypis. Centroid-Based Document Classification: Analysis and Experimental Results. 2000. http://citeseer.ist.psu.edu/rd/0%2C374513%2C1%2C0.25%2CDownload/http://citeseer.ist.psu.edu/cache/papers/cs/17801/ftp:zSzzSzftp.cs.umn.eduzSzdeptzSzuserszSzkumarzSzpkdd-cent.pdf/han00centroidbased.pdf didownload pada tanggal 23 Oktober 2007.

[8] Han, Jiawei dan Micheline Kamber. 2006. *Data Mining: Concepts and Techniques*. Second Edition. San Fransisco: Morgan Kaufmann Publisher.

[9] Reichel, Uwe D. and Florian Schiel. Using Morphology and Phoneme History to Improve Grapheme-to-Phoneme Conversion. http://www.phonetik.uni-muenchen.de/~reichelu/reichel_schiel_paper.ps didownload pada tanggal 1 Maret 2007.

[10] Tan, Pang-Ning et al. 2006. *Introduction to Data Mining*. Boston: Pearson Education, Inc.

[11] Taylor, Paul. Hidden Markov Models for Grapheme to Phoneme Conversion. 2005. http://mi.eng.cam.ac.uk/~pat40/eurospeech05_form_04.pdf didownload pada tanggal 1 Maret 2007.

[12]Yvon, Francois. Self-learning techniques for grapheme-to-phoneme conversion. 1994. http://citeseer.ist.psu.edu/rd/21528791%2C43928%2C1%2C0.25%2CDownload/http://citeseer.ist.psu.edu/cache/papers/cs/1985/http:zSzzSzwww-inf.enst.frzSz%7EresearchzSzpublications_eczSzyvonzSzyvon_94b.pdf/yvon94selflearning.pdf didownload pada tanggal 1 Maret 2007.